Understanding the Generalization Benefit of Model Invariance from a Data Perspective

Sicheng Zhu*, Bang An*, Furong Huang
Department of Computer Science
University of Maryland, College Park
{sczhu, bangan, furongh}@umd.edu

Abstract

Machine learning models that are developed to be invariant under certain types of data transformations have shown improved generalization in practice. However, a principled understanding of why invariance benefits generalization is limited. Given a task, there is often no principled way to select "suitable" data transformations under which model invariance guarantees better generalization. This paper understands the generalization benefit of model invariance by introducing the *sample cover induced by transformations*, i.e., a representative subset of a dataset that can approximately recover the whole dataset using transformations. For any data transformations, we provide refined generalization bounds for invariant models based on the sample cover. We also characterize the "suitability" of a set of data transformations by the *sample covering number induced by transformations*, i.e., the smallest size of its induced sample covers. We show that we could tighten the generalization bounds for "suitable" transformations that have a small sample covering number. In addition, our proposed sample covering number can be empirically evaluated and thus provides a guidance for selecting transformations to develop model invariance for better generalization. In experiments on multiple datasets, we empirically evaluate sample covering numbers for some commonly used transformations and show that the smaller sample covering number for a set of transformations (e.g. the 3D-view transformation) indicates a smaller gap between the test and training error for invariant models, which verifies our propositions.

1 Introduction

Invariance is ubiquitous in many real-world problems. For instance, categorical classification of visual objects is invariant to slight viewpoint changes [16, 2, 19], text understanding is invariant to synonymous substitution and minor typos [40, 28, 21]. Intuitively, models capturing the underlying invariance exhibit improved generalization in practice [18, 12, 37, 13, 34, 13]. Such generalization benefit is especially crucial when the data are scarce as in some medical tasks [35], or when the task requires more data than usual as in cases of distribution shift [29] and adversarial attack [31, 36, 5].

A commonly accepted intuition attributes the generalization benefit of model invariance to the reduced model complexity, especially the reduced sensitivity to spurious features. However, a principled understanding of why model invariance helps generalization remains elusive, thus leaving many open questions. Since model invariance may come at a cost (e.g., compromised accuracy, increase computational overhead), given a task, how should we choose among various data transformations under which model invariance guarantees better generalization? If existing data transformations are not good enough for a given task, what is the guiding principle to find new ones? The lack of a principled understanding limits better exploitation of model invariance to further improve

* Equal contribution.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

generalization. In addition, since identifying instructive generalization bound is a central topic in machine learning, we may expect to tighten existing generalization bounds by additionally considering the data-dependent model invariance property.

The many faces of data transformations and model classes pose significant challenges to a principled understanding of model invariance's generalization benefit. To address this, [2, 4, 26, 32, 30] characterize the input space and show that certain data transformations equivalently shrink the input space for invariant models, which then simplify the input and improves generalization. From another perspective, [17, 23] directly characterize the function space and show that the volume of the invariant model class is reduced, which then simplifies the learning problem and improves generalization. These understandings provide valuable insights, yet they may suffer from the high-dimensionality of input data or require the model invariance to be obtained exclusively via feature averaging. Some certain assumptions on data transformations (e.g., finiteness, group structure with certain measure) also make these understandings less applicable to more general data transformations.

In this paper, we understand the generalization benefit of model invariance by characterizing sample-dependent properties of data transformations and then provide practical guidance for selecting suitable data transformations. Our work motivates exploring more suitable data transformations on given datasets to better improve invariant models' generalization performance.

1.1 Summary of Contributions

Data-perspective understanding. (Section 3) We characterize the sample-dependent properties of data transformations by introducing the notion of sample cover induced by data transformations. We derive refined model complexity bounds for invariant models based on the new notion. The model complexity results then lead to refined generalization bounds which capture the model invariance property on the data and help understand the generalization benefit of model invariance. We also propose an algorithm to estimate the sample covering numbers in practice.

Data-driven selection of data transformations. (Section 5) To guide efficient data transformation selections in practice, we propose to use the sample covering number induced by data transformations as a suitability measurement to estimate the generalization benefit for invariant models. This measurement is data-driven, model-agnostic, and generally applicable to any data transformations. We also discuss its limitations and empirical mitigations.

Empirical evaluations. (Section 6) We empirically estimate the sample covering number for some commonly used data transformations on four image datasets, including CIFAR-10 and ShapeNet (a 3D-dataset).The 3D-view transformation renders a smaller empirical sample covering number than others by a large margin on ShapeNet, while cropping is the most favorable among others. Then, we do invariant model training under those data transformations by data augmentation and invariance loss penalization to evaluate the actual generalization benefit for invariant models. Results show a clear correlation between smaller sample covering numbers induced by data transformations and better generalization benefit for invariant models.

2 Preliminaries

Data transformations. We refer to the data transformation as a function mapping from the input space $\mathcal{X} \to \mathcal{X}$, and data transformation*s* as a set of such functions. Unless otherwise specified, we do not assume data transformations to have group structures since many non-invertible transformations (e.g., cropping) do not fit into a group structure directly. For a set of data transformations $\mathcal{G} = \{g : \mathcal{X} \to \mathcal{X}\}$ and a data point $\boldsymbol{x} \in \mathcal{X}$, we overload the notion of orbit in group theory and denote by $\mathcal{G}(\boldsymbol{x})$ the orbit of $\boldsymbol{x}$ defined as follows. The *orbit* $\mathcal{G}(\boldsymbol{x})$ of a data point $\boldsymbol{x}$ with respect to data transformations $\mathcal{G}$ is the collection of outputs after any transformation $g \in \mathcal{G}$ on $\boldsymbol{x}$: $\mathcal{G}(\boldsymbol{x}) = \{g(\boldsymbol{x}) \in \mathcal{X} : g \in \mathcal{G}\}$.

Model invariance. Let $\mathcal{D}$ be the underlying data distribution and $\text{supp}(\mathcal{D})$ be its support. A model $h : \mathcal{X} \to \mathcal{Y}$ is said to be *invariant* under data transformations $\mathcal{G}$ on $\mathcal{D}$ if $h(g(\boldsymbol{x})) = h(\boldsymbol{x})$ for any $\boldsymbol{x} \in \text{supp}(\mathcal{D})$ and any $g \in \mathcal{G}$. We refer to a class of invariant models as the *$\mathcal{G}$-invariant model class*.

Complexity measurements. *Covering number* and *Rademacher complexity* [25] are two commonly used complexity measurements for model classes (including neural networks [6]) that can provide uniform generalization bound. The covering number can also be directly used to upper bound the Rademacher complexity via Dudley's entropy integral theorem [15, 24].

Covering number. Let $(\mathcal{F}, d)$ be a (pseudo)metric space with some (pseudo)metric[1] d. An *ϵ-cover* of a set $\mathcal{H} \subseteq \mathcal{F}$ is defined as a subset $\widehat{\mathcal{H}} \subseteq \mathcal{H}$ such that for any $h \in \mathcal{H}$, there exists $\widehat{h} \in \widehat{\mathcal{H}}$ such that $d(h, \widehat{h}) \leq \epsilon$. The covering number $N(\mathcal{H}, \epsilon, d)$ is defined as the minimum cardinality of an ϵ-cover (among all ϵ-covers) of $\mathcal{H}$. In this paper, we use the concept of covering number both for measuring model class complexities and for defining the sample covering number on datasets.

Empirical Rademacher complexity. Let $\mathcal{H}$ be a class of functions $h : \mathcal{X} \to \mathbb{R}$. Given a sample $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$, the *empirical Rademacher complexity* of model class $\mathcal{H}$ is defined as: $\mathfrak{R}_{\mathcal{S}}(\mathcal{H}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \sigma_i h(\boldsymbol{x}_i)\right]$ where $\boldsymbol{\sigma} = [\sigma_1, ..., \sigma_n]^\top$ is the vector of i.i.d. Rademacher random variables, each uniformly chosen from $\{-1, 1\}$.

Generalization error and gap. Let $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$ be a sample drawn i.i.d. from some data distribution $\mathcal{D}$, and $\mathcal{H}$ be a model class. Given a loss function $\ell : \mathbb{R} \to [0, 1]$, for a $h \in \mathcal{H}$, we define the *empirical error* as $R_{\mathcal{S}}(h) = \frac{1}{n}\sum_{i=1}^n \ell(h(\boldsymbol{x}_i), y_i)$, the *generalization error* as $R(h) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell(h(\boldsymbol{x}), y)]$, and the *generalization gap* as $R(h) - R_{\mathcal{S}}(h)$.

3 Generalization Benefit of Model Invariance

In this section, we understand the generalization benefit of model invariance by refining generalization bounds for invariant models. Instead of directly characterizing the entire input space or the function space of invariant models, we first characterize the data transformations' effect on *samples* in a model-agnostic way, and then establish connections with the generalization of invariant models. Albeit less elegant, this analysis is more applicable for general data transformations regardless of how model invariance is obtained and is more tractable for empirical evaluation.

We first introduce the notion of *sample cover induced by data transformations*, which is a representative subset of examples induced by the data transformations. The intuition is that invariant models may have consistent behavior on a data point and its approximation in the sample cover so that we only need to consider the invariant models' behavior on the sample cover.

The definition of sample cover relies on a pseudometric induced by the data transformations $\mathcal{G}$. Let $\|\cdot\|$ be any norm on the input space $\mathcal{X}$ (which is used in defining the line integral and the model's Lipschitz constant). Given the set of transformations $\mathcal{G}$, we define the pseudometric[2]

$$\rho_{\mathcal{G}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \inf_{\gamma\in\Gamma(\boldsymbol{x}_1,\boldsymbol{x}_2)} \int_\gamma c(\boldsymbol{r})ds, \quad \text{where } c(\boldsymbol{r}) = \begin{cases} 0, & \text{if } \boldsymbol{r} \text{ is in some orbit} \\ 1, & \text{otherwise} \end{cases}$$

where $ds = \|d\boldsymbol{r}\|$ and Γ denotes the set of all paths in $\mathcal{X}$ from $\boldsymbol{x}_1$ to $\boldsymbol{x}_2$. The pseudometric is essentially calculating the line integral along the shortest (if achievable) path γ with weighted cost.

Definition 3.1 (Sample cover induced by data transformations)**.** Let $(\mathcal{X}, \rho_{\mathcal{G}})$ be a pseudometric space and $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$ be a sample of size n. An *ϵ-sample cover* $\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}$ of the sample $\mathcal{S}$ induced by data transformations $\mathcal{G}$ at resolution ϵ is defined as a representative subset of the sample $\mathcal{S}$ such that for any $\boldsymbol{x} \in \mathcal{S}$, there exists $\widehat{\boldsymbol{x}} \in \widehat{\mathcal{S}}_{\mathcal{G},\epsilon}$ such that $\rho_{\mathcal{G}}(\boldsymbol{x}, \widehat{\boldsymbol{x}}) \leq \epsilon$.

Definition 3.2 (Sample covering number induced by data transformations)**.** The *sample covering number* $N(\epsilon, \mathcal{S}, \rho_{\mathcal{G}})$ induced by data transformations $\mathcal{G}$ is defined as the minimum cardinality of an ϵ-sample cover:

$$N(\epsilon, \mathcal{S}, \rho_{\mathcal{G}}) = \min\{|\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}| : \widehat{\mathcal{S}}_{\mathcal{G},\epsilon} \text{ is an } \epsilon\text{-sample cover of } \mathcal{S}\}. \tag{3.1}$$

The sample cover induced by some data transformations $\mathcal{G}$ enables us to compute the $\mathcal{G}$-invariant models' complexity by considering only the models' behaviour on the sample cover. Moreover, the sample covering number induced by $\mathcal{G}$ directly leads to the following model class complexity bound which holds for any $\mathcal{G}$-invariant model class.

[1] A pseudometric is a metric if and only if it separates distinct points, namely $d(x, y) > 0$ for any $x \neq y$.

[2] Note that $\rho_{\mathcal{G}}$ is not a metric since it allows $\rho_{\mathcal{G}}(x, y) = 0$ for $x \neq y$.

Proposition 3.3. Let $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$ be a sample of size n. Let $\mathcal{H}$ be a model class that has bounded output and is invariant to $\mathcal{G}$. Then the following Rademacher upper-bound for $\mathcal{H}$ holds up to some constant.

$$\mathfrak{R}_{\mathcal{S}}(\mathcal{H}) \leq 24\sqrt{\frac{N(0, \mathcal{S}, \rho_{\mathcal{G}})}{n}}. \tag{3.2}$$

Proposition 3.3 shows that any $\mathcal{G}$-invariant model class has a bounded empirical Rademacher complexity in terms of the sample covering number $N(0, \mathcal{S}, \rho_{\mathcal{G}})$ induced by $\mathcal{G}$. Smaller sample covering number thus gives tighter model class complexity bound for invariant models.

However, Proposition 3.3 is informative only when the data transformations yield $N(0, \mathcal{S}, \rho_{\mathcal{G}}) \ll n$ on the sample $\mathcal{S}$, which means we can exactly recover all data points in $\mathcal{S}$ using a small-sized subset and data transformations $\mathcal{G}$. This condition is typically not satisfied by data transformations which only generate orbits with measure zero (with respect to data measure) at most data points. For example, the rotation transformations on CIFAR-10 do not satisfy this condition, since no two images in CIFAR-10 are rotated versions of each other. To relax this restriction on data transformations to better understand the generalization benefit brought by more commonly used data transformations (e.g., rotation), we further consider specific model class properties which expand the data point orbits so as to get more refined results. We study the Lipschitz model class in Section 3.1. A more restrictive yet sharper and more interpretable result for linear models appears in Appendix A.

3.1 Refined Model Complexity for Lipschitz Models

This subsection shows the refined model complexity result brought by invariance against any data transformations, assuming Lipschitz model class. The following theorem provides a refined covering number result in terms of the original covering number of the model class.

Theorem 3.4 (Refined model covering number of $\mathcal{G}$-invariant $\mathcal{H}^{\mathsf{Lip}}$)**.** Let $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$ be a sample of size n. Let $\mathcal{H}^{\mathsf{Lip}}$ be a model class such that every $h \in \mathcal{H}^{\mathsf{Lip}}$ is κ-Lipschitz with respect to $\|\cdot\|$ (which is used in defining the sample cover) and invariant to $\mathcal{G}$. Then

$$N\big(\tau, \mathcal{H}^{\mathsf{Lip}}, L_2(\mathbb{P}_{\mathcal{S}})\big) \leq \inf_{\epsilon \geq 0, \widehat{\mathcal{S}}_{\mathcal{G},\epsilon}} N\Big(\tau - \kappa\epsilon\sqrt{1 - \frac{|\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}|}{n}}, \mathcal{H}^{\mathsf{Lip}}, L_2(\mathbb{P}_{\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}})\Big), \tag{3.3}$$

where the (pseudo)metrics[3] are defined as $\|h - g\|_{L_2(\mathbb{P}_{\mathcal{S}})} = \Big(\sum_{\boldsymbol{x}\in\mathcal{S}} \frac{1}{n}\big(h(\boldsymbol{x}) - g(\boldsymbol{x})\big)^2\Big)^{\frac{1}{2}}$ and $\|h - g\|_{L_2(\mathbb{P}_{\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}})} = \Big(\sum_{\boldsymbol{x}\in\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}} \frac{p(\boldsymbol{x})}{n}\big(h(\boldsymbol{x}) - g(\boldsymbol{x})\big)^2\Big)^{\frac{1}{2}}, \forall h, g \in \mathcal{H}^{\mathsf{Lip}}$.

Remark 3.5. Theorem 3.4 shows that the previous covering number of $\mathcal{H}^{\mathsf{Lip}}$ evaluated at the sample $\mathcal{S}$ is upper-bounded by a new covering number of $\mathcal{H}^{\mathsf{Lip}}$ evaluated at any sample cover $\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}$, under a modified metric and at the cost of an additional error term that depends on ϵ. The equality trivially holds by taking $\widehat{\mathcal{S}}_{\mathcal{G},\epsilon} = \mathcal{S}$, and by searching over all sample covers with different resolution ϵ it is possible to tighten the covering number for the invariant Lipschitz model class. Previous results on the covering number require $N\big(\tau, \mathcal{H}, L_2(\mathbb{P}_{\mathcal{C}})\big)/n \to 0$ as $n \to \infty$ so as to be non-vacuous (e.g., see [39]). Covering numbers linearly depending on n is thus usually replaced by some heavy dependence on the input data dimension, which may lead to loose result when the data dimension is high (e.g., images). Now with the refined result in Theorem 3.4, covering numbers depending on n can potentially be replaced by the one depending on sample covering numbers $N\big(\tau, \mathcal{H}, L_2(\mathbb{P}_{\mathcal{C}})\big)$, thus steering away from the dependence on data dimension.

Theorem 3.4 directly leads to a refined empirical Rademacher complexity bound in terms of the covering number of $\mathcal{H}$ evaluated at the sample cover. The following proposition is a weaker but more convenient version. We can set $\epsilon \to 0$ as $n \to \infty$ to further suppress the additional error term on large samples.

[3] The term $p(\boldsymbol{x})$ in defining the function norm $L_2(\mathbb{P}_{\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}})$ can be viewed as the probability mass at $\boldsymbol{x}$ which equals to the number of examples that $\boldsymbol{x}$ covers. The formal definition of $p(\boldsymbol{x})$ is relegated to the Appendix B

Proposition 3.6 (Refined Rademacher complexity of $\mathcal{G}$-invariant $\mathcal{H}^{\mathsf{Lip}}$)**.** Let $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$ be a sample of size n. Let $\mathcal{H}^{\mathsf{Lip}}$ be a model class such that every $h \in \mathcal{H}^{\mathsf{Lip}}$ is κ-Lipschitz with respect to $\|\cdot\|$ and invariant to $\mathcal{G}$. Given any $\epsilon > 0$, $\alpha > 0$, let $\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}$ be an ϵ-cover of $\mathcal{S}$. Then

$$\mathfrak{R}_{\mathcal{S}}(\mathcal{H}) \leq 4\kappa\epsilon\sqrt{1 - \frac{|\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}|}{n}} + 4\alpha + 12\int_{\alpha}^{\infty}\sqrt{\frac{\log N\left(\tau, \mathcal{H}, L_2(\mathbb{P}_{\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}})\right)}{n}}d\tau. \tag{3.4}$$

3.2 Model-Invariance-Dependent Generalization Bounds

This subsection applies the refined invariant model complexity results to present model-invariance-dependent generalization bounds. Those generalization bounds identify model invariance properties and provide a justification for seeking appropriate types of model invariance depending on the dataset.

Invariant models. Note that for invariant model classes, we can directly plug the corresponding refined model complexity results (Proposition 3.3 or Proposition 3.6 or Theorem A.1) into the standard generalization bound (see Theorem B.2) to give the refined generalization bound for trained models. For invariant model class selection, we can further make the generalization bound uniform over all choices of data transformations and use the structural risk minimization [25] to control the possible increase of empirical error. Nevertheless, assuming model invariance may be too restrictive for post-hoc analysis of unknown or non(-strictly)-invariant model classes (e.g., standard models trained with data augmentations).

Other models. For trained models that are not assumed to be invariant, it is also possible to give model-invariance-dependent generalization bounds after composition with certain losses. For example, losses that are averaged or adversarially perturbed over the orbit of an input point are invariant to data transformations with group structures [23, 17]. We now use adversarial loss as an example to give the model-invariance-dependent generalization bound for general models.

Given a set of data transformations $\mathcal{G}$, we define the associated adversarial loss as $\widetilde{l}_{\mathcal{G}}\left(h(x), y\right) = \max_{x' \in \mathcal{G}(x)} l\left(h(x), y\right)$, where an input is perturbed within $\mathcal{G}$ to maximize the original loss. Given a set of different types of group-structured data transformations $\{\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, ..., \mathcal{G}^{(L)}\}$, we also construct the combined data transformations set indexed by k as follows: fix an arbitrary listing of the power set of $[L]$ and index the subsets $\mathcal{I}_k$ by k. For any $k \in [2^L]$, construct the group-structured[4] data transformation set $\mathcal{G}_k$ as the direct product of some group-structured data transformations chosen by $\mathcal{I}_k$: $\mathcal{G}_k = \Pi_{i \in \mathcal{I}_k}\mathcal{G}^{(i)}$. Then, the proposition below uses the structural risk minimization framework [25] to give the model-invariance-dependent generalization bound.

Proposition 3.7. Given a sample $\mathcal{S}$, a model $h \in \mathcal{H}$ trained on $\mathcal{S}$, and a set of different types of group-structured data transformations $\{\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, ..., \mathcal{G}^{(L)}\}$, let $\mathcal{G}_k$ be the combined data transformation set constructed as above and indexed by k. Let $\widetilde{l}_{\mathcal{G}_k}$ be the associated adversarial loss. Then the following generalization bound holds for any $k \in [2^L]$:

$$R(h) \leq \frac{1}{n}\sum_{i=1}^{n}\widetilde{l}_{\mathcal{G}_k}(h(x_i), y_i) + 4\mathfrak{R}_{\mathcal{S}}(\widetilde{l}_{\mathcal{G}_k} \circ \mathcal{H}) + \sqrt{\frac{\log k}{n}} + 3\sqrt{\frac{\log\frac{4}{\delta}}{2n}} \tag{3.5}$$

Proposition 3.7 suggests a potential trade-off: a data transformation set $\mathcal{G}_k$ may reduce the invariant model complexity $\mathfrak{R}_{\mathcal{S}}(\widetilde{l}_{\mathcal{G}_k} \circ \mathcal{H})$, but tends to increase the empirical error $\sum_{i=1}^n \widetilde{l}_{\mathcal{G}_k}(h(x_i), y_i)$ by requiring a more strict transformation-invariant output. Thus, if we can find an appropriate value of k where the the empirical error remains small, then the model h benefit from a tightened generalization bound. As shown in our experiment, models trained with data augmentation tend to have more invariant output even though the strict model invariance is not guaranteed, and this result shows their generalization benefit.

The difficulty in applying Proposition 3.7 in practice is that the model complexity with adversarial loss may be hard to compute for general models and data transformations, although Lipschitz loss l and data transformations preserve the models' Lipschitzness after composition $\mathfrak{R}_{\mathcal{S}}(\widetilde{l}_{\mathcal{G}_k} \circ \mathcal{H})$. Therefore, we discuss more practical implications of the sample covering result in the next section.

[4]Note that direct product preserves the group structure.

Algorithm 1 Empirical Estimation of Sample Covering Number

Input: Sample $\mathcal{S}$, a set of transformations $\mathcal{G}$, resolution ϵ

Output: $\widehat{N}(\epsilon, \mathcal{S}, \rho_{\mathcal{G}})$, an empirical estimation of sample covering number $N(\epsilon, \mathcal{S}, \rho_{\mathcal{G}})$

Algorithm:

Step 1: Compute (or approximate) the direct orbit distance between any two examples $\boldsymbol{x}_s, \boldsymbol{x}_t \in \mathcal{S}$.

$$d_{\mathcal{G}}(\boldsymbol{x}_s, \boldsymbol{x}_t) = \|\mathcal{G}(\boldsymbol{x}_s) - \mathcal{G}(\boldsymbol{x}_t)\| = \min_{g_1, g_2 \in \mathcal{G}} \|g_1(\boldsymbol{x}_s) - g_2(\boldsymbol{x}_t)\|.$$

Step2: Compute the $\rho_{\mathcal{G}}$ distance between any two examples $\boldsymbol{x}_s, \boldsymbol{x}_t \in \mathcal{S}$ by solving a shortest path problem:

$$\rho_{\mathcal{G}}(\boldsymbol{x}_s, \boldsymbol{x}_t) = \min_{path(s,t)} \sum_{(i,j) \in path(s,t)} d_{\mathcal{G}}(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

Step3: Construct the pairwise distance matrix $\mathbf{D} \leftarrow [\rho_{\mathcal{G}}(\boldsymbol{x}_s, \boldsymbol{x}_t)]_{s,t}$ and approximate the sample covering number $\widehat{N}(\epsilon, \mathcal{S}, \rho_{\mathcal{G}}) \leftarrow$ `Distance2SampleCoveringNum`($\mathbf{D}$,ϵ) using Algorithm 2.

4 Empirical Estimation of Sample Covering Numbers

Exactly computing the sample covering number is intractable in general since computing the transformation induced metric can be difficult and finding the smallest sample cover is NP-hard. To address this problem, we propose an algorithm to empirically estimate the sample covering number. Algorithm 1 outlines the algorithm and the full detail appears in Algorithm 2 in Appendix D. We describe the algorithm steps below and then discuss the algorithmic challenges.

Step 1. Compute (or approximate) the direct orbit distance between any two examples in $\mathcal{S}$. The direct orbit distance between any two examples $\boldsymbol{x}_s, \boldsymbol{x}_t \in \mathcal{S}$ is

$$d_{\mathcal{G}}(\boldsymbol{x}_s, \boldsymbol{x}_t) = \|\mathcal{G}(\boldsymbol{x}_s) - \mathcal{G}(\boldsymbol{x}_t)\| = \min_{g_1, g_2 \in \mathcal{G}} \|g_1(\boldsymbol{x}_s) - g_2(\boldsymbol{x}_t)\|,$$

which can be exactly computed for finite transformations (e.g. flipping) with complexity $O(|\mathcal{G}|^2)$), or can be approximated for continuous transformations (e.g. rotation) via optimization or sampling.

Step 2. Compute the $\rho_{\mathcal{G}}$ distance between any two examples in $\mathcal{S}$. Given results in step 1, Computing the $\rho_{\mathcal{G}}$ distance between any two examples can be formulated as a shortest path problem on a complete graph, where each node represents an example and the cost of each edge is the direct orbit distance computed in step 1. Appendix D provides the associated shortest path problem. For each pair of examples, this problem can be solved by standard shortest path algorithms (e.g. Dijkstra's algorithm) in polynomial time (e.g. $O(n^3)$).

Step 3. Construct the pairwise distance matrix $\mathbf{D} \leftarrow [\rho_{\mathcal{G}}(\boldsymbol{x}_s, \boldsymbol{x}_t)]_{s,t}$ and approximate the sample covering number. This step can be formulated as a set cover problem where each example $\boldsymbol{x}$ forms a subset of $\mathcal{S}$ in which each element's $\rho_{\mathcal{G}}$ distance to $\boldsymbol{x}$ is less than or equal to ϵ. Our goal is to find a minimum number of those subset such that their union contains $\mathcal{S}$. This problem is known to be NP-hard but admits polynomial time approximations. In experiments, we use modified k-medoids [27] clustering method to find the approximation of $N(\epsilon, \mathcal{S}, \rho_{\mathcal{G}})$ (see Algorithm 2).

Algorithmic challenge. The computational bottleneck of this algorithm is in step 2, especially when the dataset is large. Instead, we can evaluate sample covering numbers on a subset sampled i.i.d. from the training set in practice. Assume that the training set is also sampled i.i.d. from the underlying population distribution. When we consider the ratio between the sample covering number induced by a certain group of data transformations and that without any data transformations, the estimation of it based on a subset sampled i.i.d. from the population distribution is an unbiased estimation. Appendix E empirically show that by changing the size of this subset, the order of sample covering numbers of different data transformations preserves.

5 Data-driven Selection of Data Transformations

The pool of candidate data transformations on a given dataset may be infinitely large. To make full use of the generalization benefit of invariant models, we usually make selections based on expensive cross-validations due to the absence of a-priori guidance. In Section 3, we show that an invariant model is likely to generalize well if the corresponding data transformations induce a small sample covering number. Therefore, we propose to use sample covering number as an empirical suitability measurement to guide the data transformations selection and discuss its advantages, limitations, and empirical mitigations.

Suitability measurement. For a set of data transformations $\mathcal{G}$ and a dataset $\mathcal{S}$, we use the sample covering number $N(\epsilon, \mathcal{S}, \rho_{\mathcal{G}})$ induced by $\mathcal{G}$ as the suitability measurement of $\mathcal{G}$ on $\mathcal{S}$ in terms of the generalization benefit for invariant models.

Advantages. This suitability measurement only depends on datasets and data transformations and is thus model-training-free. It provides a-priori guidance for the data transformations selection without the need of expensive model training. Such an efficient measurement may fuel the exploration of new types of data transformations to improve invariant models' generalization performance. Another advantage of this suitability measurement is that it is generally applicable to all types of data transformations, including continuous and non-invertible ones, and provides a uniform comparison.

Limitation and mitigation: empirical error. Unlike the uniform generalization bound presented in Proposition 3.7, using only the sample covering number as the measurement only considers the generalization gap and ignores the model invariance's effect on the empirical error. Inappropriate model invariance may drastically increase the empirical error and override the benefit of reduced generalization gap. To mitigate this limitation, we provide two necessary conditions to keep the empirical error low. First, the data transformations should preserve the underlying ground-truth labeling. We use domain knowledge to ensure this condition. Second, the model class should be rich enough to contain a low-error invariant hypothesis. In our experiment, neural networks achieving a low invariance training error suffices to meeting this condition.

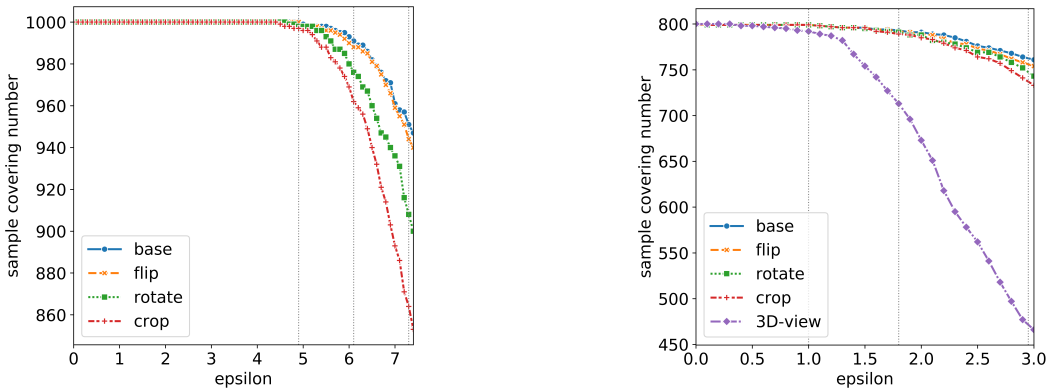
Limitation and mitigation: Lipschitz constant. Theorem 3.4 suggests that the generalization benefit enjoyed by invariant models depends on models' Lipschitz constant and can be negated if enforcing invariance leads to significantly larger Lipschitz constant. Since the suitability measurement is model-agnostic, predicting and offsetting the Lipschitz constant change seems impossible without actually training a model. To mitigate this limitation, we use the fact that we are doing classification task and use the label information to empirically offset the potential Lipschitz constant increase for selecting data transformations. We use the minimum inter-class distance change after considering data transformations to address the Lipschitz constant change and use it to normalize the sample covering number (see Appendix E.5.2).

6 Experiments

In this section, we empirically verify whether our proposed sample covering number induced by data transformations correlates the actual generalization benefit for invariant models. Specifically, we first measure the sample covering number induced by different types of data transformations using the estimation Algorithm 1. Then, we investigate the actual generalization benefit for models invariant to those transformations in classification tasks and analyze the correlation[5].

Datasets. We conduct experiments on two public datasets, CIFAR-10 [22] and ShapeNet [9], in this section. ShapeNet is a large-scale 3D data repository which enables us to do more complex data transformations (e.g., change of 3D-view) beyond commonly used 2D geometric transformations. In our experiment, we use a subset of ShapeNet which contains 10 categories. [11] provide, for each 3D object, a set of 24 multi-view pre-rendered images, and we use them to approximate the random perturbation of 3D-view. In Appendix E.5.1, we also conduct experiments on CIFAR-100 and Restricted ImageNet.

[5] Code is available at https://github.com/bangann/understanding-invariance.



(a) CIFAR-10 (b) ShapeNet

Figure 1: Estimated sample covering numbers induced by different data transformations at different resolutions ϵ. "base" indicates no transformation. Note that as ϵ increases, it starts to exceed the L_2 distance between some images and thus some images get covered by others without doing any transformation. Three vertical dashed lines indicate the maximum resolution ϵ at which the "base" yields a certain sample covering number, and from left to right they are $100\%n$, $99\%n$, $95\%n$.

Data transformations. We choose some commonly used geometric transformations with typical settings (which are assumed to be label-preserving). On CIFAR-10, we choose *flipping*, *cropping*, and *rotation*. On ShapeNet, we choose *flipping*, *cropping*, *rotation*, and additionally the *3D-view* change. We choose the same type of data transformations with the same setting during estimating the sample covering number and evaluating the generalization benefit. Appendix E provides more details on settings and results.

6.1 Estimation of Sample Covering Numbers

We apply Algorithm 1 to estimate the sample covering number induced by different data transformations. For efficiency, we randomly choose 1000 training images on CIFAR-10 and randomly choose 800 training images on ShapeNet. We choose L_2 norm as the input space norm used in Definition 3.2 which measures the distance between different images. For continuous data transformations, we do random sampling to approximate the orbit of an data point. Appendix E provides more estimation results with different sample sizes.

Sample covering number. Figure 1 illustrates the estimated sample covering numbers induced by different transformations at different resolutions ϵ. As the resolution ϵ increases, the sample covering number $N(\epsilon, \mathcal{S}, \rho_{\mathcal{G}})$ induced by any data transformation starts to decrease, indicating a smaller-sized sample cover needed to cover the entire dataset. Meanwhile, different transformations behave differently. On CIFAR-10, cropping induces the smallest sample covering number. On ShapeNet, 3D-view transformation induces the smallest sample covering number and the gap is significant. Notably, 3D-view transformation can cover the dataset ($n = 1000$) with 80% of the data while keeping ϵ below the minimum distance between any two images. According to our propositions, data transformations that are able to induce a smaller sample covering number would benefit more for the corresponding invariant models. Therefore, Figure 1 indicates that models should generalize well if it is invariant to 3D-view transformation on ShapeNet or cropping on CIFAR-10.

6.2 Evaluation of Generalization Benefit

To see how data transformations benefit generalization and to verify whether sample covering number is a good indicator of the generalization benefit, we learn corresponding invariant models for classification tasks and analyze the actual generalization benefit. We use ResNet18 [20] on both datasets and discuss the influence of model class's implicit bias in Appendix E. A simple method to learn invariant models is to do data augmentation. The augmented loss function is $\mathcal{L}_{aug}(\boldsymbol{x}) = \mathcal{L}(f(g(\boldsymbol{x})))$, where $f(\cdot)$ denotes the model and $g(\boldsymbol{x})$ denotes a randomly sampled

Model	$n = 100$		$n = 1000$		$n = all$	
	acc (%)	gap	acc (%)	gap	acc (%)	gap
Base	41.05 ± 0.52	58.95 ± 0.52	68.62 ± 0.90	31.38 ± 0.90	85.43 ± 0.35	14.57 ± 0.35
Flip	44.19 ± 0.74	55.81 ± 0.74	75.12 ± 0.20	24.88 ± 0.20	89.67 ± 0.24	10.33 ± 0.24
Rotate	47.02 ± 0.46	52.93 ± 0.51	76.07 ± 0.28	23.92 ± 0.27	89.91 ± 0.13	10.05 ± 0.16
Crop	$\mathbf{50.47 \pm 0.48}$	$\mathbf{49.53 \pm 0.48}$	$\mathbf{81.84 \pm 0.12}$	$\mathbf{18.15 \pm 0.11}$	$\mathbf{92.52 \pm 0.08}$	$\mathbf{7.48 \pm 0.08}$

Table 1: Classification accuracy and generalization gap (the difference between training and test accuracy) for ResNet18 on CIFAR-10. n denotes the sample size per class.

Model	$n = 100$		$n = 1000$		$n = all$	
	acc (%)	gap	acc (%)	gap	acc (%)	gap
Base	67.75 ± 2.02	32.25 ± 2.02	83.33 ± 0.38	16.67 ± 0.38	91.81 ± 0.22	8.18 ± 0.22
Flip	69.75 ± 1.55	30.25 ± 1.55	84.24 ± 0.30	15.76 ± 0.30	92.07 ± 0.20	7.92 ± 0.20
Rotate	70.25 ± 1.19	29.50 ± 1.15	83.93 ± 0.38	15.94 ± 0.35	91.85 ± 0.20	8.03 ± 0.26
Crop	74.88 ± 1.03	23.53 ± 1.30	86.13 ± 0.39	13.75 ± 0.32	92.64 ± 0.12	7.17 ± 0.19
3D-View	$\mathbf{78.13 \pm 1.31}$	$\mathbf{14.94 \pm 1.76}$	$\mathbf{88.79 \pm 0.34}$	$\mathbf{8.38 \pm 0.79}$	$\mathbf{94.38 \pm 0.08}$	$\mathbf{3.09 \pm 0.10}$

Table 2: Classification accuracy and generalization gap (the difference between training and test accuracy) for ResNet18 on ShapeNet. n denotes the sample size per class.

example in $\boldsymbol{x}$'s orbit induced by transformation $\mathcal{G}$. We use this method on CIFAR-10 and ShapeNet and show results in table 1 and 2.

Sample covering number correlates well with generalization benefit. We use the generalization gap (the gap between training accuracy and test accuracy) to measure actual generalization benefit. Compared with the baseline, invariant models show an improved reduced generalization gap and also improved test accuracy. On CIFAR-10, cropping-invariant model shows the smallest generalization gap and the highest accuracy. On ShapeNet, the model that is invariant to 3D-view changes shows the smallest generalization gap and the highest accuracy, especially when the training data size is small. By comparing results in figure 1 and table 1-2, we observe a clear correlation between the smaller sample covering number and better generalization benefit. This verifies our proposition — invariance to more suitable data transformations gives the model more generalization benefit.

Model invariance indeed improves after learning. We conduct another experiment on ShapeNet with 3D-view transformation to verify that the improved generalization is indeed brought by model invariance. Instead of using data augmentation, we train models with an invariance regularization loss similar to [38]: $\mathcal{L}_{reg} = \mathcal{L}_{cls} + \mathcal{L}_{inv} = \mathcal{L}(f(\boldsymbol{x})) + \lambda \mathrm{KL}(f(\boldsymbol{x}), f(g(\boldsymbol{x})))$. Specifically, in addition to minimize the classification loss on the original image, we penalize the model by minimizing the KL divergence between model outputs of the original image and the transformed one. By

λ	train acc (%)	test acc (%)	gap	$\mathcal{L}_{inv}$	$\mathcal{A}_{inv}(\%)$
0	99.99 ± 0.01	91.81 ± 0.22	8.19 ± 0.22	0.0548 ± 0.0028	62.0 ± 0.6
0.01	99.98 ± 0.00	92.77 ± 0.16	7.21 ± 0.16	0.0290 ± 0.0029	74.78 ± 1.61
0.1	99.99 ± 0.00	93.87 ± 0.19	6.11 ± 0.19	0.0152 ± 0.0003	83.12 ± 0.50
0.3	99.98 ± 0.00	94.23 ± 0.11	5.76 ± 0.11	0.0121 ± 0.0003	85.10 ± 0.20
1	99.58 ± 0.04	94.68 ± 0.09	4.90 ± 0.09	0.0095 ± 0.0001	86.94 ± 0.08
3	97.74 ± 0.19	94.48 ± 0.19	3.26 ± 0.09	0.0060 ± 0.0003	88.15 ± 0.18
10	95.67 ± 0.26	93.56 ± 0.29	2.11 ± 0.04	0.0037 ± 0.0002	89.20 ± 0.16
100	92.89 ± 0.25	91.85 ± 0.26	1.03 ± 0.03	0.0018 ± 0.0001	89.82 ± 0.10

Table 3: Evaluation of ResNet18 on ShapeNet under 3D-view transformations. $\mathcal{L}_{inv}$ denotes the test invariance loss. $\mathcal{A}_{inv}$ denotes the test consistency accuracy (indicating whether model's prediction is unchanged after data transformation) under worst-case transformations.

adjusting λ, we can control the invariance extent. At test time, we use an empirical approximation of $\mathcal{L}_{inv}(\boldsymbol{x}) = \mathbb{E}_{g_1,g_2\in\mathcal{G}}[\mathrm{KL}(f(g_1(\boldsymbol{x})), f(g_2(\boldsymbol{x})))]$ to evaluate model invariance under transformation $\mathcal{G}$. In Table 3, as we increase the invariance penalty by increasing λ, invariant models show smaller generalization gap. Moreover, the decreased invariance loss and increased consistency accuracy verify that the model invariance indeed improves after learning, and thus suggests that the generalization benefit is brought by model invariance.

7 Related Work

Input space perspective. The most related line of work is the perspective that characterizes the input space. [3, 4] show that the invariant representations can equivalently reduce the input dimension for downstream tasks and thus significantly reduce the model complexity (which is exponentially dependent on input dimensions) of downstream linear models. [32] factorizes the input space into the product of a base space and a finite set of data transformations. Since the covering number needed to cover the base space is smaller, they show that it leads to a reduced generalization bound for large margin classifiers. [30] also factorizes the input space using finite data transformations and show the reduced covering number results in improved generalization bound for invariant models. Compared with these works, our work tries to cover the sample instead of the input space which circumvents the exponential dependence on input dimensions and enables empirical evaluation.

Function space perspective. Another line of work directly characterizes the function space of invariant models. [23] uses PAC-Bayes to show the reduction of generalization upper bound when the model class is symmetrized to be invariant. [17] analyzes the function space under feature averaging operator and shows the first strict generalization gap (instead of upper bound) via a linear model. So far this line of work only considers the model invariance obtained exclusively via feature averaging.

Others. We mention some work that addresses the model invariance but does not focus on understanding the benefit. [1] mentions that the VC-dimension of an invariant model cannot be larger than its counterpart. [8] characterizes the general functional representations of invariant probability distributions and neural network structures that implement them. [10] uses group theory to show the benefit of learning with data-augmented loss.

8 Conclusion

This paper understands the generalization benefit of model invariance from a data perspective. By introducing sample cover induced by data transformations, we characterize the properties of data transformations and relates it to the improved generalization bound for invariant models. Our theoretical and empirical results motivate the exploration of more suitable data transformations on a given task and develop the corresponding invariant models for better generalization performance. One limitation of our work is that our model-agnostic way does not take into account the implicit bias of certain model classes, and thus cannot exactly predict the generalization benefit an invariant model would have. Besides, we do not have an extensive discussion on balancing the trade-off between improved generalization gap and reduced training accuracy. In future work, we would try to consider the implicit bias of certain model classes to better understand the benefit of model invariance.

Acknowledgements

This work is supported by a startup fund from the Department of Computer Science of the University of Maryland, National Science Foundation IIS CRII Award, DOD-DARPA-Defense Advanced Research Projects Agency Guaranteeing AI Robustness against Deception (GARD), Air Force Material Command, and Adobe, Capital One and JP Morgan faculty fellowships.

References

[1] Yaser S. Abu-Mostafa. Hints and the VC Dimension. *Neural Comput.*, 5(2):278–288, March 1993. Place: Cambridge, MA, USA Publisher: MIT Press.

[2] Fabio Anselmi, Joel Z. Leibo, Lorenzo Rosasco, Jim Mutch, Andrea Tacchetti, and Tomaso Poggio. Unsupervised learning of invariant representations in hierarchical architectures. *arXiv:1311.4158 [cs]*, Mar 2014. arXiv: 1311.4158.

[3] Fabio Anselmi, Joel Z. Leibo, Lorenzo Rosasco, Jim Mutch, Andrea Tacchetti, and Tomaso Poggio. Unsupervised Learning of Invariant Representations in Hierarchical Architectures. *arXiv:1311.4158 [cs]*, March 2014. arXiv: 1311.4158.

[4] Fabio Anselmi, Lorenzo Rosasco, and Tomaso Poggio. On Invariance and Selectivity in Representation Learning. *arXiv:1503.05938 [cs]*, March 2015. arXiv: 1503.05938.

[5] Pranjal Awasthi, Natalie Frank, and Mehryar Mohri. Adversarial Learning Guarantees for Linear Hypotheses and Neural Networks. *arXiv:2004.13617 [cs, stat]*, April 2020. arXiv: 2004.13617.

[6] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems 30*, pages 6240–6249. Curran Associates, Inc., 2017.

[7] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[8] Benjamin Bloem-Reddy and Yee Whye Teh. Probabilistic symmetries and invariant neural networks. *arXiv:1901.06082 [cs, stat]*, September 2020. arXiv: 1901.06082.

[9] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[10] Shuxiao Chen, Edgar Dobriban, and Jane Lee. A Group-Theoretic Framework for Data Augmentation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21321–21333. Curran Associates, Inc., 2020.

[11] Christopher B. Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, volume 9912 of *Lecture Notes in Computer Science*, pages 628–644. Springer, 2016.

[12] Taco Cohen and Max Welling. Group Equivariant Convolutional Networks. In *International Conference on Machine Learning*, pages 2990–2999. PMLR, June 2016. ISSN: 1938-7228.

[13] Taco S. Cohen, Mario Geiger, and Maurice Weiler. A general theory of equivariant cnns on homogeneous spaces. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily B. Fox, and Roman Garnett, editors, *NeurIPS*, page 9142–9153, 2019.

[14] Terrance Devries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *CoRR*, abs/1708.04552, 2017.

[15] R. M. Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

[16] Shimon Edelman. Class similarity and viewpoint invariance in the recognition of 3d objects. *Biological Cybernetics*, 72(3):207–220, Feb 1995.

[17] Bryn Elesedy and Sheheryar Zaidi. Provably Strict Generalisation Benefit for Equivariant Models. *arXiv:2102.10333 [cs, stat]*, February 2021. arXiv: 2102.10333.

[18] Robert Gens and Pedro M. Domingos. Deep symmetry networks. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *NIPS*, page 2537–2545, 2014.

[19] Yena Han, Gemma Roig, Gad Geiger, and Tomaso Poggio. Scale and translation-invariance for novel objects in human vision. *Scientific Reports*, 10(1):1411, Jan 2020.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[21] Erik Jones, Robin Jia, Aditi Raghunathan, and Percy Liang. Robust encodings: A framework for combating adversarial typos. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 2752–2765. Association for Computational Linguistics, Jul 2020.

[22] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[23] Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the Benefits of Invariance in Neural Networks. *arXiv:2005.00178 [cs, stat]*, April 2020. arXiv: 2005.00178.

[24] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012.

[25] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT press, 2018.

[26] Youssef Mroueh, Stephen Voinea, and Tomaso A Poggio. Learning with group invariant features: A kernel perspective. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[27] Hae-Sang Park and Chi-Hyuck Jun. A simple and fast algorithm for k-medoids clustering. *Expert systems with applications*, 36(2):3336–3341, 2009.

[28] Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. Combating adversarial misspellings with robust word recognition. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *ACL (1)*, page 5582–5591. Association for Computational Linguistics, 2019.

[29] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

[30] Akiyoshi Sannai and Masaaki Imaizumi. Improved Generalization Bound of Group Invariant / Equivariant Deep Networks via Quotient Feature Space. *arXiv:1910.06552 [cs, stat]*, March 2020. arXiv: 1910.06552.

[31] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially Robust Generalization Requires More Data. In *Advances in Neural Information Processing Systems 31*, pages 5014–5026. Curran Associates, Inc., 2018.

[32] Jure Sokolic, Raja Giryes, Guillermo Sapiro, and Miguel R. D. Rodrigues. Generalization Error of Invariant Classifiers. *arXiv:1610.04574 [cs, stat]*, July 2017. arXiv: 1610.04574.

[33] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Mądry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

[34] Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco S Cohen. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[35] Marysia Winkels and Taco S. Cohen. 3d g-cnns for pulmonary nodule detection. *CoRR*, abs/1804.04656, 2018.

[36] Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, page 7085–7094. PMLR, Jun 2019.

[37] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *NIPS*, page 3391–3401, 2017.

[38] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Learning Representations*, 2019.

[39] Tong Zhang. Covering number bounds of certain regularized linear function classes. *J. Mach. Learn. Res.*, 2:527–550, Mar 2002.

[40] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *NIPS*, page 649–657, 2015.